

The logo for IRIA (Institut National de Recherche en Informatique et en Automatique) is displayed in a stylized, bold, white font against a dark, textured background.

CENTRE DE ROCQUENCOURT

# Rapports de Recherche

N° 440

## **ANALYSE DISCRIMINANTE : MÉTHODE DU TYPE PLUS PROCHES VOISINS UTILISANT UN PRÉTRAITEMENT DES DONNÉES**

**Fabrice BONNEAU  
Jean-Marie PROTH**

Institut National  
de Recherche  
en Informatique  
et en Automatique

Domaine de Voluceau  
Rocquencourt  
BP 105

78153 Le Chesnay Cedex  
France

Tél. (3) 954 90 20

**Septembre 1985**

**ANALYSE DISCRIMINANTE :**

**METHODE DU TYPE PLUS PROCHES VOISINS  
UTILISANT UN PRETRAITEMENT DES DONNEES**

Fabrice BONNEAU \*

Jean-Marie PROTH \*\*

\* : INRIA, Projet SAGEP, Domaine de Voluceau, ROCQUENCOURT B.P. 105,  
78153 LE CHESNAY CEDEX

\*\* : INRIA, Projet SAGEP, Château du Montet, 54500 VANDOEUVRE



## RESUME

Dans la suite, nous proposons une méthode d'analyse discriminante essayant de concilier les avantages d'une méthode géométrique et d'une méthode par voisinage.

Dans la première partie nous présentons un algorithme de prétraitement des données, permettant d'obtenir des régions homogènes et connexes.

Nous proposons ensuite une règle de décision, s'appuyant sur cette nouvelle structure des données, tantôt géométrique tantôt du type plus proche voisin.

Nous décrivons enfin la procédure particulière "plus proche voisin" que nous utilisons.

## ABSTRACT

In the following, we propose a discriminant analysis method which tries to conciliate advantages of a geometrical method and of a nearest neighbour method.

In the first part, we present a data preprocessing algorithm in order to find homogeneous and connex regions.

Getting from the new datas structure, we propose a decision rule, sometimes geometrical, sometimes using a particular procedure of nearest neighbours.

## I - RAPPELS

Nous nous intéressons ici à l'aspect prédictif de l'analyse discriminante et non à l'aspect descriptif. Ainsi, nous supposons connu l'ensemble des prédicteurs que l'on a pu obtenir par sélection pas à pas, par analyse factorielle discriminante ou par n'importe quelle autre méthode.

### I - 1 LE MODELE

Soit donc :

- E : une population d'individus, éventuellement infinie, chacun d'eux étant défini par p variables quantitatives (prédicteurs).
- Une variable qualitative w à k modalités connues a priori sur un sous-ensemble fini B de E et que l'on désire prédire sur l'ensemble E tout entier. B est appelé population de Base. La variable w partitionne donc l'ensemble B en m classes  $A_1, \dots, A_m$ .

On appelle  $Cr = \{x \in E, w(x) \in r\}$  pour  $r = 1, \dots, m$

On aura alors  $Ar = Cr \cap B$

Une règle d'affectation D sera une application de E-B dans  $(1, \dots, m)$  telle que  $D(x) = r$  représentera la décision d'affecter x à  $Cr$ .

### I - 2 METHODES GEOMETRIQUES

On désigne par  $Gr_r$   $r = 1, \dots, m$ , les centres de gravité respectifs des ensembles  $Ar$ . On se donne une métrique M sur E de sorte que

$$d(x, y) = \sqrt{(x-y)^T M (x-y)}$$

définisse une distance dans E.

On peut proposer comme règle d'affectation  $D(x) = r_0$  où l'on a :

$$d(x, G_{r_0}) = \inf_{r=1, \dots, m} \{ d(x, Gr_r) \}$$

Remarque : La métrique peut dépendre de r.

#### CHOIX DE LA METRIQUE

Il est toujours possible de prendre la distance usuelle mais on démontre que le meilleur choix est

$$dr(x, Gr_r) = \sqrt{(x-Gr_r)^T Mr(x-Gr_r)}$$

ou

$$Mr = (\det(Vr))^{1/p} \cdot Vr^{-1}$$

V est la matrice variance-covariance associée à  $Ar$ .

Cette distance tient compte de la forme des ensembles  $A_r$ . En effet l'ensemble

$$\{x / dr(x, Gr) = R_0\}$$

est, dans le cas de la dimension 2, une ellipse de centre  $Gr$ , de grand axe, l'axe principal d'inertie.

Ainsi si

$$R_0 = \max_{y \in A_r} \{dr(y, Gr)\} \quad \text{alors } dr(x, Gr) < R_0$$

peut être considéré comme une inéquation de  $A_r$ .

La distance  $dr$  s'appelle distance de Mahalanobis associée à la population  $A_r$ .

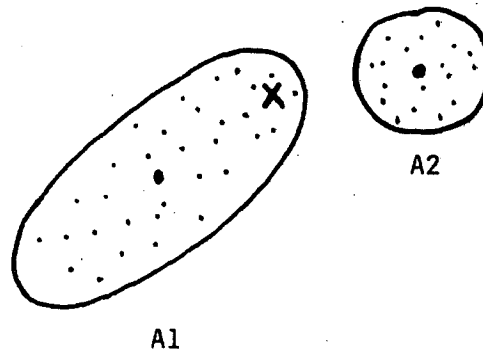
#### REMARQUE

Si l'on souhaite effectuer des calculs plus simples, il est possible de prendre

$$M_r = (\text{var}(x_1) \dots \text{var}(x_p))^{1/p} \text{diag}(1/\text{var}(x_i), i = 1, p)$$

On démontre que cette métrique est la meilleure métrique associée à une matrice diagonale répondant à la question.

#### EXEMPLE

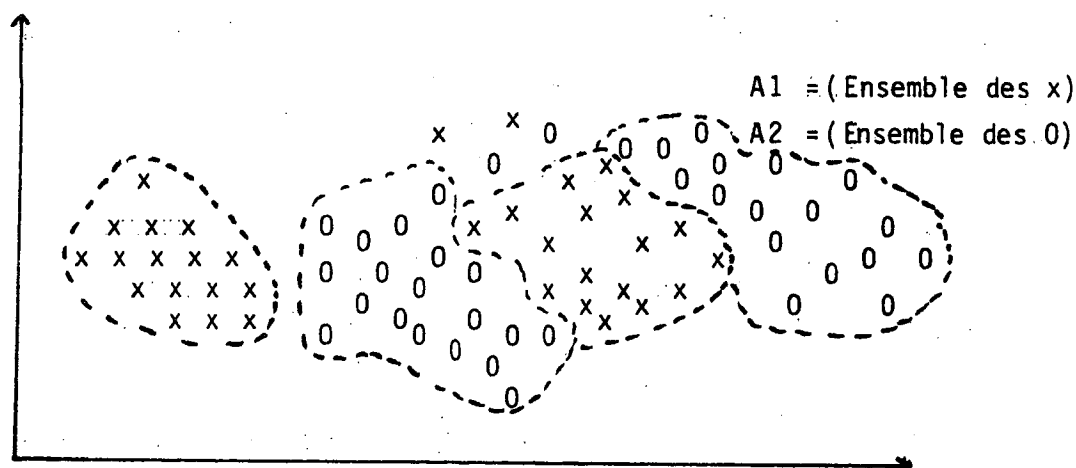


Avec la distance de Mahalanobis,  $x$  sera affecté à sa classe naturelle ( $A_1$ ), avec la distance euclidienne il sera affecté à  $A_2$ .

#### INCONVENIENTS DE LA METHODE

Cette méthode sera efficace si les ensembles  $A_r$  sont géométriquement discernables, c'est-à-dire si ils sont connexes et non mélangés. En effet, l'ensemble des prédicteurs peut-être bien choisi sans pour autant garantir ces deux conditions.

Soit par exemple pour  $q = 2$ ,  $p = 2$  :



Les ensembles  $A_1$  et  $A_2$  sont tous deux formés de deux régions géométriques et d'une zone mélangée : l'affectation géométrique sera catastrophique.

### I - 3 METHODE PAR VOISINAGE

Lorsqu'on désire affecter  $x \in E$ , on calcule ses  $k$  plus proches voisins dans  $B$  où  $k$  est un entier donné à priori. On affecte ensuite à la majorité : si les  $k$  plus proches voisins de  $x$  sont  $V = \{x_1, x_2, \dots, x_k\}$  on décide d'affecter  $x$  à  $r_0$  si

$$\text{card}(V \cap A_{r_0}) = \max_{r=1, m} \{\text{card}(V \cap A_r)\}$$

La distance choisie pour calculer les plus proches voisins sera la plupart du temps la distance euclidienne, même si l'on peut imaginer prendre :

$$M = (\det(V))^{1/p_V - 1},$$

où  $V$  est la matrice de variance-covariance totale, afin de pondérer correctement les variables.

Cette méthode est souvent efficace quant à la précision de l'affectation obtenue : c'est une méthode locale qui ne dépend guère de la disposition géométrique des ensembles  $A_r$  et qui repose sur un principe empirique : "deux individus très proches ont de grandes chances d'appartenir au même  $C_r$ ".

Malheureusement, elle est difficilement applicable telle quelle dans la pratique :

- La population de base peut être nombreuse (plus de 500) et le temps de calcul est alors très long. On peut trouver dans la littérature plusieurs

algorithmes de recherche des k plus proches voisins ([1],[2],[3],[4]) qui ont des performances comparables.

- Etant une méthode locale, l'affectation k.p.p.v ne tient pas compte de la forme globale des classes à discriminer. Elle peut être mauvaise lorsque c'est "la tendance générale" de la répartition qui importe.

- Il peut être difficile de trouver le meilleur k dans une méthode k.p.p.v.

## II - PRINCIPE DE LA METHODE PROPOSEE

Notre but n'est pas seulement de proposer un algorithme rapide k.p.p.v, mais d'essayer de surmonter en partie l'ensemble des inconvénients énoncés plus haut.

Pour cela on effectue deux opérations :

- Un prétraitement de l'ensemble B afin de déceler comment sont géométriquement répartis les ensembles  $A_r$ . B sera ainsi partitionné en 1+1 classes, 1 régions connexes et homogènes du point de vue des  $A_r$  et la zone hétérogène restante que l'on appellera zone trouble. Ce prétraitement a lieu une fois pour toutes.
- Une règle d'affectation mixte tantôt directe (géométrique), tantôt par voisinage, les plus proches voisins étant recherchés non pas sur tout l'ensemble mais sur une "fenêtre" dont le rayon est la distance au centre de gravité de la région la plus proche.

Le principe est donc de mener à bien le mieux possible la première étape (quitte à y passer beaucoup de temps), la 2ème étape étant ainsi instantanée.

## III - PRETRAITEMENT

### III - 1 QUELQUES DEFINITIONS

#### III-1.1. ENSEMBLES CONNEXES DANS E

Par analogie avec la topologie nous appellerons boule de E de rayon s et de centre x l'ensemble formé par x et ses s plus proches voisins, et ensemble connexe tout sous-ensemble C de E vérifiant la propriété suivante, notée P :

P) Pour tout  $(x,y)$  de  $C^2$  il existe une suite de boules  $B_0, \dots, B_l$ , incluses dans  $C$ , telles que pour tout  $i$  de 0 à  $l-1$

$$B_i \cap B_{i+1} \neq \emptyset \text{ avec } x \in B_0 \text{ et } y \in B_l$$

Intuitivement, cette notion est tout à fait analogue à la notion de connexité en topologie et nous nous appuierons sur cette analogie pour la construction d'un algorithme.

### III-1.2. REGION DE E

On dira qu'un sous-ensemble  $B$  de  $E$  est d'épaisseur au moins  $s_0$  si il existe une boule de rayon  $s_0$  incluse dans  $B$ . On dira qu'il est de taille au moins  $s_1$  si son cardinal est au moins égal à  $s_1$ .

On appellera REGION  $(s_0, s_1)$  de  $E$  tout sous-ensemble connexe, d'épaisseur au moins  $s_0$  et de taille au moins  $s_1$ . On a bien sur :

$$0 \leq s_0 < s_1$$

#### REMARQUE

Les régions  $(0,1)$  sont les connexes de  $E$ .

### III-1.3. ELEMENTS CARACTERISTIQUES D'UNE REGION R

On définit tout d'abord les grandeurs habituelles :

- Centre de gravité :  $G$ , moyenne des éléments de la région.
- Métrique de Mahalanobis :  $M = (\det(V))^{1/P} V^{-1}$  où  $V$  = matrice de variance-covariance  $d_R(x,y) = t(x-y) M(x-y)$  est la distance de Mahalanobis associée à  $R$ .
- rayon maximal :  $r_{\max} = \max_{x \in R} \{d_R(x, G)\}$

On peut également définir :

- rayon minimal :  $r_{\min} = \min_{y \notin R} \{d_R(y, G)\}$
- Ellipse circonscrite :  $\{x \in R^P / d_R(x, G) = r_{\max}\}$
- Ellipse inscrite :  $\{x \in R^P / d_R(x, G) = r_{\min}\}$



On peut remarquer que toutes ces notions peuvent être définies indépendamment de la notion de région mais elles ne deviennent significatives que si l'ensemble considéré vérifie les propriétés d'une région ou tout au moins des propriétés voisines. Si par exemple on considère l'ensemble formé par 2 disques disjoints, l'ellipse inscrite est réduite du point  $G$  qui n'appartient d'ailleurs même pas à l'ensemble.

#### REMARQUE IMPORTANTE

La notion de région telle que nous l'avons définie est beaucoup moins forte que la convexité, ces régions pouvant avoir toutes les formes possibles pourvu qu'elles restent connexes. Ceci pour deux raisons :

- Algorithmiquement il est très difficile de reconnaître les zones convexes.
- La convexité est une hypothèse agréable mais exigeante et l'on risque d'obtenir des régions de cardinal bien trop faible.

### **III - 2. PARTITIONNEMENT EN REGIONS**

#### III-2.1. DEFINITION

Soit  $A$  un sous-ensemble d'un ensemble fini  $E$ . On dira que  $R_1, \dots, R_l, Z$  est un partitionnement en régions  $(s_0, s_1)$  si les ensembles  $R_1, \dots, R_l, Z$  vérifient :

- i) pour tout  $i < l$ ,  $R_i$  est une région de taille au moins  $s_1$  et d'épaisseur au moins  $s_0$ ,
- ii) pour tout couple  $(i, j)$  l'ensemble  $R_i \cup R_j$  n'est plus une région,
- iii) le sous-ensemble  $Z$  est tel que :
  - pour tout  $x$  de  $Z$  et pour tout  $i$ , l'ensemble  $R_i \cup \{x\}$  n'est plus une région,
  - il n'existe pas de région  $(s_0, s_1)$  incluse dans  $Z$ .

#### III-2.2. THEOREME 1

Pour  $s_0$  et  $s_1$  donnés et pour tout sous-ensemble  $A$  de  $E$ , il existe un et un seul partitionnement en régions  $(s_0, s_1)$  de  $A$ .

### DEMONSTRATION

. Cas  $s_0=0$ ,  $s_1=1$  (cas connexe, il n'y pas de zone trouble).

Le théorème est l'équivalent du théorème de topologie d'existence et d'unicité de la décomposition en composantes connexes.

#### LEMME 1

Si  $C_1, C_2$  sont deux ensembles connexes de  $E$  et s'il existe une boule  $B$  incluse dans  $C_1 \cup C_2$  et telle que  $B \cap C_1 \neq \emptyset$  et  $B \cap C_2 \neq \emptyset$ , alors  $C_1 \cup C_2$  est connexe.

#### Démonstration

La démonstration est évidente : pour rejoindre un point  $x$  de  $C_1$  et un point  $y$  de  $C_2$  il suffit de transiter par la boule  $B$ . On joint  $x$  à  $x_1$  par une suite de boules où  $x_1 \in C_1 \cap B$ , on choisit un point  $x_2$  de  $C_2 \cap B$  et on joint  $x_2$  à  $y$  par une suite de boules : la réunion des deux suites obtenues permet de joindre  $x$  à  $y$ .

#### LEMME 2

Si  $C_1, C_2$  est un partitionnement en deux ensembles connexes de  $C_1 \cup C_2$  alors pour tout  $A_1, A_2$  connexes tels que  $A_1 \subset C_1$ ,  $A_2 \subset C_2$ ,  $A_1, A_2$  est un partitionnement de  $A_1 \cup A_2$ .

#### Démonstration

On suppose que  $A_1 \cup A_2$  est connexe. Alors il existe nécessairement une boule de  $A_1 \cup A_2$  contenant à la fois des éléments de  $A_1$  et de  $A_2$ . Il suffit d'appliquer le Lemme 1.

#### Existence

On le démontre par récurrence sur le nombre d'éléments de  $A$ . C'est évident pour le singleton ( $n=1$ ). Supposons donc la propriété vraie pour les ensembles à  $n$  éléments et soit  $A$  un ensemble à  $n+1$  éléments. Soit  $x$  de  $A$ , il existe un partitionnement de  $A - \{x\}$  noté  $C_1, \dots, C_k$ . Si pour tout  $j, C_j \cup \{x\}$  n'est pas connexe alors  $C_1, \dots, C_k, \{x\}$  est un partitionnement de  $A$ . Sinon, soit par exemple  $C_1, \dots, C_l$  tels  $\{x\} \cup C_1 \cup \dots \cup C_l$  soit connexe et  $\forall j, l+1 \leq j \leq k$ ,  $\{x\} \cup C_1 \cup \dots \cup C_l \cup C_j$  n'est plus connexe alors  $\{x\} \cup C_1 \cup \dots \cup C_l, C_{l+1}, \dots, C_k$  est un partitionnement de  $A$ .

### Unicité

Supposons tout d'abord que A est connexe. Il faut alors montrer que le seul partitionnement de A est A lui-même.

Considérons donc  $A_1, A_2, \dots, A_n$  un partitionnement de A. Soit x de  $A_i$  et y de  $A_j$ ,  $i \neq j$ . Alors, comme A est connexe, il existe une suite de boules de A "reliant" x à y. Soit B la première boule qui n'est pas incluse dans  $A_i$  et  $x_{i0}$  son centre. On suppose que  $x_{i0} \in A_{i0}$ . Soit alors  $B_0(x_{i0}, r_0)$ , la boule de centre  $x_{i0}$  et de rayon  $r_0$  où :

$$r_0 = \min\{r / \text{il existe } z \in A - A_{i0} \text{ et } z \in B\}.$$

Cette boule existe nécessairement puisque  $B \not\subset A_{i0}$ , que  $i=i_0$  ou pas. Supposons que  $z \in A_{j_0}$  où  $j_0 \neq i_0$ . Il suffit d'appliquer le Lemme 1 à  $A_{i0}$  et  $A_{j_0}$ . Donc  $A_{i0} \cup A_{j_0}$  est connexe ce qui contredit la propriété ii).

Soit maintenant A quelconque et  $C_1, \dots, C_n$  et  $D_1, \dots, D_m$  deux partitionnements de A. Soit  $F_i = C_1 \cap D_i, i=1, m$ . On fait un partitionnement de chaque  $F_i$ . D'après le Lemme 2 on obtient ainsi un partitionnement de  $C_1$ , ce qui n'est possible, d'après ce qui précède, que si tous les ensembles sauf un sont vides. Il vaut  $C_1$ . On a donc montré qu'il existe j tel que  $C_1 = D_j$ . On raisonne de même avec  $C_2, \dots, C_n$ .

. Cas  $s_0, s_1$  quelconques

### Existence

C'est évident : on fait un partitionnement en ensembles connexes puis on obtient Z par la réunion de tous les ensembles connexes n'étant pas  $(s_0, s_1)$ . On vérifie que Z possède les propriétés souhaitées par unicité du partitionnement en ensembles connexes.

### Unicité

Soit  $R_1, \dots, R_n, Z$  et  $S_1, \dots, S_m, T$  deux partitionnements en région  $(s_0, s_1)$  de A. Alors, si l'on partitionne Z et T en ensembles connexes, on sait, d'après ce qui précède, que les partitionnements  $R_1, \dots, R_n, Z_1, \dots, Z_k$  et  $S_1, \dots, S_m, T_1, \dots, T_l$  sont identiques, mais on ne peut avoir  $Z_i = S_j$  car  $Z_i$  n'est pas une région  $(s_0, s_1)$  et, de même, on ne peut avoir  $T_i = R_j$ . On a donc finalement  $Z = Z_1 \cup \dots \cup Z_k = T_1 \cup \dots \cup T_l = t$  et les  $R_i, T_j$  égaux deux à deux.

C.Q.F.D.

### III-2.3. CONCLUSION

L'intérêt d'un tel partitionnement est d'obtenir le découpage de  $Ar$  en régions bien définies géométriquement et la possibilité d'utiliser l'ensemble de notions définies précédemment dans un algorithme d'affectation.

Le but de l'algorithme de prétraitement est donc non seulement de trouver le bon partitionnement, mais également de calculer pour chaque région obtenue ses grandeurs caractéristiques.

## III - 3 ALGORITHME DE PRETRAITEMENT

### III-3.1. PRINCIPE DE L'ALGORITHME

On se donne  $(s_0, s_1)$  et une classe  $Ar$  que l'on veut partitionner en régions  $(s_0, s_1)$  :

- on initialise  $Z$  à  $Ar$ ,
- on parcourt l'ensemble  $Ar$  dans l'ordre du fichier, soit un point  $x$  de  $Ar$
- on calcule la plus grande boule de centre  $x, Bx$ , incluse dans  $Ar$ . Plusieurs cas peuvent alors se présenter :
  - .  $\text{ray}(Bx) < s_0$  et il n'existe pas de région  $R$  déjà formée telle que  $Bx \cap R \neq \emptyset$  ; on passe à l'élément de  $Ar$  suivant,  $x$  restant dans  $Z$ .
  - .  $\text{ray}(Bx) \geq s_0$  et il n'existe pas de région  $R$  déjà formée telle que  $Bx \cap R \neq \emptyset$  ; une région  $R = Bx, (s_0, \text{card}(Bx))$ , est formée.
  - . il existe des régions  $R_1, \dots, R_m$  déjà formées telles que  $Bx \cap R_j \neq \emptyset$  ; une région  $R = Bx \cup R_1 \cup \dots \cup R_m$  est alors formée,
- on parcourt  $Ar$  jusqu'à ce que le partitionnement se stabilise,
- on remet dans  $Z$  toutes les régions de cardinal inférieur à  $s_1$ .

### III-3.2. THEOREME 2

Le partitionnement de  $Ar$  obtenu est l'unique partitionnement en régions  $(s_0, s_1)$  de  $Ar$ .

#### Démonstration

On vérifie que les sous-ensembles sont bien des régions  $(s_0, s_1)$ . Ceci est pratiquement évident et repose sur deux propositions élémentaires :

toute boule est connexe et l'union de connexes dont l'intersection n'est pas vide est connexe.

Les autres points de la définition se vérifient eux aussi aisément.

### III-3.3. PRETRAITEMENT : UTILISATION CONCRETE DE L'ALGORITHME

(Dans toute la suite p.p.v. signifie "plus proche(s) voisin(s)")

On effectue le partitionnement pour tous les  $A_r$  après avoir choisi les seuils  $s_0$  et  $s_1$  (ces seuils dépendent de la taille du fichier de base, du nombre de variables, et plus intuitivement du nombre de p.p.v. que l'on considèrera dans l'algorithme d'affectation. Nous y reviendrons plus tard).

On calcule ensuite les grandeurs caractéristiques définies en III-1.3.

On obtient ainsi :

- ntr régions ( $s_0, s_1$ ) de  $E$  ( $R_1, R_2, \dots, R_{ntr}$ ) avec leur marquage  $W$  :  $W(i) = r$  où  $A_r$  est le sur-ensemble de  $R_i$  ( $R_i$  est une région où les éléments ont pour modalité  $r$ ),
- on associe à chaque région son centre de gravité  $G_i$ , sa métrique  $M_i$  (ou  $d_i$ ), ses rayons : maximal  $R_{MAXi}$  et minimal  $R_{MINi}$

#### Quelques remarques :

- La métrique choisie sera soit la métrique usuelle, soit la métrique associée à l'inverse de la matrice de variance co-variance totale (ce dernier choix étant nécessaire si l'on n'est pas sûr de disposer de variables indépendantes et de même importance prédictive).
- l'algorithme peut être long mais pratiquement 2 ou 3 itérations suffisent pour un ensemble  $A_r$ .
- Il se peut que pour un  $A_r$  on ne trouve aucune région : on peut alors choisir  $s_0$  plus petit pour amorcer l'algorithme.
- Il est bien entendu intéressant d'avoir des régions de grande épaisseur et de cardinal élevé : l'algorithme d'affectation n'en sera que plus performant.

## IV - ALGORITHME D'AFFECTATION

### IV-1. PRINCIPE GENERAL

Comme cela a été précisé au début nous voulons, d'une part un algorithme rapide, et d'autre part une affectation tenant compte de la forme des classes  $Ar$  c'est-à-dire de leur découpage en régions et de la forme de ces régions. La méthode proposée est une méthode du type p.p.v. mais précédée de trois tests d'appartenance directe à une région. De plus, les plus proches voisins ne sont calculés que sur un sous-ensemble de  $B$  et on calcule les  $k$  plus proches voisins de  $k = k_{min}$  à  $k_{max}$ , un critère étant ensuite choisi qui fournit le meilleur  $k$ . L'affectation n'a lieu que si le critère maximum obtenu est supérieur à un certain seuil de tolérance ; dans le cas contraire on préférera affecter à la région la plus proche.

### IV-2. DESCRIPTION DE L'ALGORITHME

Soit deux entiers  $k_{min}$ ,  $k_{max}$  avec  $k_{min} < k_{max}$  et un nombre  $tol$  ( $0 \leq tol < 1$ ).

Soit  $X_{new}$  un point de  $E$  que l'on désire affecter à une modalité  $r$  donc à un ensemble  $Cr$ .

On note :  
-  $d_i(X_{new}, G_i)$ , la distance de Mahalanobis entre  $X_{new}$  et  $G_i$ ,  
-  $D(X_{new}, x)$ , la distance usuelle.

(1) On calcule les distances  $d_i(X_{new}, G_i)$  et on appelle

$R_0 = \min_i \{d_i(X_{new}, G_i)\}$  et  $i_0$  l'indice de la région pour laquelle ce min est atteint.

(2) Test 1 :

Si il existe  $i_s$  tel que  $d_{i_s}(X_{new}, G_{i_s}) < R_{MINis}$  alors on affecte directement  $X_{new}$  à  $W(i_s)$ .

(3) On considère  $\{i_1, \dots, i_t\}$  l'ensemble des  $i$  tels que

$$d_i(X_{new}, G_i) < R_0 + R_{MAXi},$$

et 
$$Z_0 = \{x \in Z / D(X_{new}, x) < R_0\}$$

(4) Test 2

Si  $t = 0$  et  $\text{card}\{Z_0\} < k_{min}$  : on affecte  $X_{new}$  à  $W(i_0)$

(5) Si  $t > 0$  on considère la fenêtre

$$F = Z_0 \cup \{x \in R_{i_1} / d_{i_1}(X_{new}, x) < R_0\} \cup \dots$$

$$\dots \cup \{x \in R_{i_t} / d_{i_t}(X_{new}, x) < R_0\}$$

l'ensemble des points de B à une distance de  $X_{new}$  inférieure à  $R_0$ .

- Test 3 : Si  $\text{card}(F) < k_{\min}$  on affecte  $X_{new}$  à  $W(i_0)$

- Sinon, on appelle une procédure de recherche de plus proches voisins à l'intérieur de  $F$  :  $KPPV(k_{\min}, k_{\max}, \text{tol})$  qui détermine l'affectation de  $X_{new}$ . (Voir détails de la procédure en IV-3.).

#### Commentaires

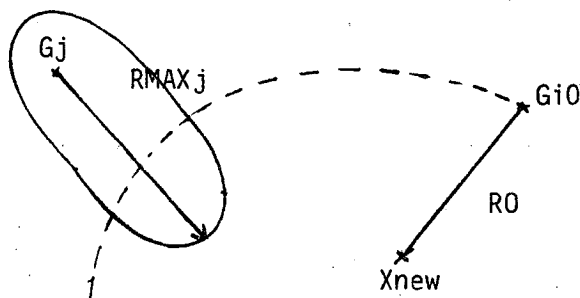
L'étape 1 consiste simplement à calculer la région la plus proche et fournit également le rayon de la fenêtre sur laquelle on calculera les P.P.V.

L'étape 2 est le test qui détermine si oui ou non le point  $X_{new}$  tombe à l'intérieur de l'ellipse inscrite.

A l'étape 3 on calcule l'ensemble des régions concernées, c'est-à-dire pouvant avoir une intersection non vide avec la fenêtre  $\{x/d(X_{new}, x) < R_0\}$  (On voit facilement que ces régions sont telles que  $d_i(X_{new}, G_i) < R_0 + R_{MAXi}$ , voir dessin).

L'étape 4 teste le nombre de régions concernées. Si celui ci est égal à 1 et si, de plus, le nombre d'éléments de la zone trouble tombant dans  $F$  est suffisamment faible, on affectera  $X_{new}$  à la classe associée à la seule région concernée.

A l'étape 5 on appelle la procédure K.P.P.V., sauf si le nombre d'éléments de la fenêtre est trop faible, auquel cas on affectera  $X_{new}$  à la classe associée à la région la plus proche.



### IV-3 LA PROCEDURE K.P.P.V.

#### IV-3.1. RAPPEL

Le principe de l'affectation du type plus proche voisin repose sur la notion intuitive suivante : si  $K$  est un entier ( $K > 0$ ) et  $V$  est la boule de centre  $X_{new}$  et de rayon  $K$  alors, on peut estimer la densité de probabilités de l'appartenance de  $X_{new}$  à  $Ar$  par :

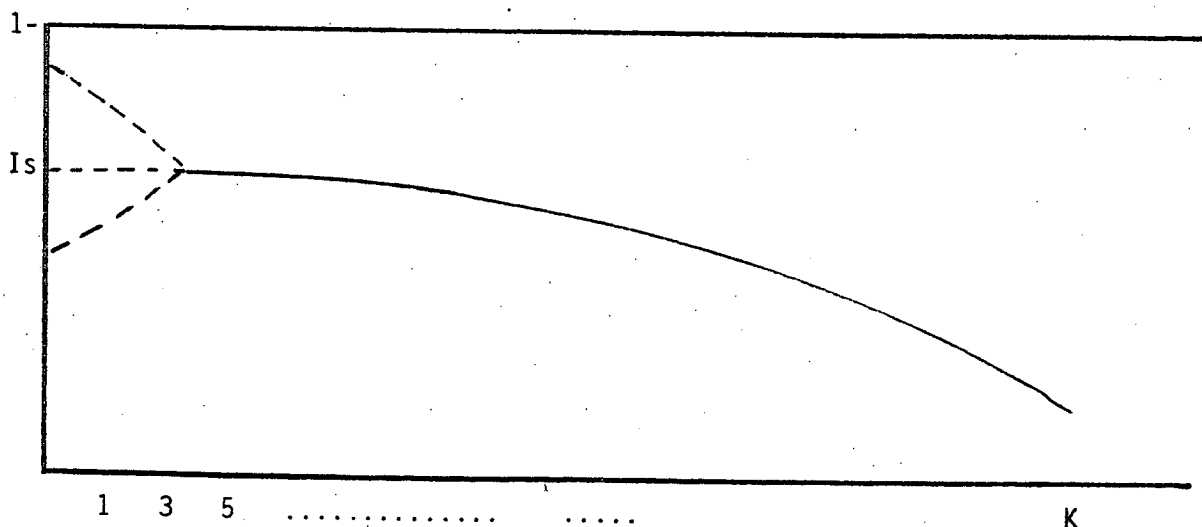
$$P(X_{new} \in Ar) = \text{card}(V \cap Ar) / K$$

qui représente la proportion d'éléments de  $Ar$  à l'intérieur du Volume  $V$ .

Dans une affectation K.P.P.V. on affectera donc  $X_{new}$  à la modalité maximisant  $\text{card}(V \cap Ar)$

#### IV-3.2. CHOIX DE K

Le choix de  $K$  est le problème le plus délicat pour l'utilisation d'un algorithme des K.P.P.V., puisqu'il dépend d'une multitude de facteurs dont l'influence est difficile à apprécier : taille du fichier, nombre de prédicteurs, nombre de classes, caractéristiques statistiques du fichier, etc... La solution empirique qui semble la plus satisfaisante est de calculer le coefficient  $Is(K)$  mesurant, pour chaque  $K$ , le pourcentage de réussite sur le fichier d'apprentissage lorsqu'on effectue une affectation K.P.P.V., et de tracer la courbe  $(K, is(K))$  : cette courbe aura en général la forme suivante :



(pour un exemple précis, voir la partie 5)



### Remarques

On ne considère que les valeurs impaires de  $K$ , les cas d'égalité des votes étant moins fréquents. (i.e. on a en général  $Is(2K) < Is(2K-1)$  et  $Is(2K) < Is(2K+1)$ ).

La forme de la courbe pour les petites valeurs de  $K$  est très variable, le maximum pouvant même être atteint pour  $K = 1$ .

Il est possible après examen de la courbe, de choisir une bonne valeur de  $K$ .

### IV-3.3. LA PROCEDURE KPPV ( $kmin, kmax, tol$ )

Dans ce paragraphe nous proposons donc une nouvelle procédure dont le principe est de calculer les  $K$  plus proches voisins de  $K = kmin$  jusqu'à  $kmax$ , ces valeurs pouvant être choisies à partir de la courbe précédente. On exige seulement :

$$1 \leq kmin < kmax$$

( $kmax$  peut être, par exemple, la valeur à partir de laquelle la courbe décroît plus rapidement,  $kmin$  peut être pris égal à 3).

La métrique utilisée peut être la distance usuelle. Mais la plupart du temps on préférera prendre pour  $d(X_{new}, x)$  la distance de Mahalanobis associée à la région à laquelle appartient  $x$  ou bien la distance usuelle si  $x$  appartient à la zone trouble. On tient compte ainsi de la forme des régions.

A chaque étape (pour chaque  $K$ ) on calcule :

- $V(K)$  l'ensemble des K.P.P.V. de  $X_{new}$
- $Maj(K) = \max \{ \text{card}(V(K) \cap A_r) \}$
- $im(K)$ , la modalité ayant obtenu la majorité  $Maj(K)$
- $P(K) = Maj(K)/K$ , estimation de la probabilité de  $X_{new}$  d'appartenir à  $C_{im}$  (i.e. d'avoir la modalité  $im$ ).

On décidera d'affecter  $X_{new}$  à  $im(K_0)$  si :

$$P(K_0) = \max [P(K)] \text{ pour } k = kmin, \dots, kmax$$

et

$$P(K_0) > tol$$

Si pour tout  $K$ ,  $P(K) < tol$ , on affectera  $X_{new}$  à la modalité associée à la région la plus proche.

### THEOREME 3

Soit  $K_0$  le premier  $K$  tel qu'on ait  $P(K) > \max(tol, kmax/(K+kmax))$ .

Alors l'affectation de  $X_{new}$  sera  $im(K_0)$ .

### Remarques

- il se peut qu'il n'y ait pas de K satisfaisant cette relation
- pour  $K_0 = K_{\max}$ ,  $\text{tol} < 0.5$ , la relation revient à dire que  $\text{im}(K_0)$  a la majorité absolue
- ce résultat inclut bien entendu le cas  $\text{Maj}(K_0) = K_0$  i.e.  $P(K_0) = 1$

### Preuve

On peut prouver que, si il existe un  $K_1 > K_0$  tel que  $P(K_1) > P(K_0)$ , alors  $\text{im}(K_1) = \text{im}(K_0)$

Supposons donc qu'il y ait un  $K > K_0$ , avec  $\text{im}(K) \neq \text{im}(K_0)$ , tel que  $P(K) > P(K_0)$ . Soit  $\text{Maj}(K) / K > \text{Maj}(K_0) / K_0$ . Le cas le plus favorable pour que l'on ait cette relation est que :

- de  $K_0+1$  jusqu'à  $K$ , la boule  $V$  ne soit remplie qu'avec des éléments de la classe  $\text{im}(K)$
- $K = k_{\max}$  (le critère n'en sera que meilleur)
- à l'étape  $K_0$ , tous les éléments qui n'appartenaient pas à  $\text{im}(K_0)$  appartenaient à  $\text{im}(K)$

$P(k_{\max})$  peut alors s'exprimer par

$$(K_0 - \text{Maj}(K_0) + k_{\max} - K_0) / k_{\max} = (k_{\max} - \text{Maj}(K_0)) / k_{\max}$$

mais  $P(K_0) > k_{\max} / (k_{\max} + K_0) \implies \text{Maj}(K_0) > (K_0 \times k_{\max}) / (K_{\max} + K_0)$

et donc  $P(k_{\max}) < (k_{\max} - (K_0 \times k_{\max}) / (k_{\max} + K_0)) / k_{\max} = 1 - K_0 / (k_{\max} + K_0)$

et donc  $P(k_{\max}) < k_{\max} / (k_{\max} + K_0) < P(K_0)$

C.Q.F.D.

### Conséquences

Ce théorème permet un gain de temps conséquent puisque, le plus souvent, on arrêtera les calculs avant  $K = k_{\max}$ .

Si l'on écrit la relation sous la forme :

$$\text{Maj}(K_0) / (K_0 - \text{Maj}(K_0)) > k_{\max} / K_0$$

ceci peut permettre de choisir  $k_{\max}$  en fonction de ce que l'on souhaite avoir comme type d'affectation pour les petites valeurs de  $K$ .

### Exemple

Supposons que l'on estime que, pour  $K=4$ , une majorité à 3 contre 1 ne doive pas toujours être suffisante. Il faudra prendre, pour que cette majorité ait une chance d'être battue,  $k_{\max} > 4 \times 3 / 1 = 12$ .

Par contre si l'on estime que pour  $K=5$ , une majorité à 4 contre 1 est assez forte pour que l'on ne souhaite pas aller plus loin, on prendra  $k_{\max} < 5 \times 4 = 20$ .

#### THEOREME 4

Si à l'étape  $K$  on a la relation  $(Maj+k_{\max}-K)/k_{\max} < tol$  alors l'affectation de  $X_{\text{new}}$  sera la région la plus proche.

#### Preuve

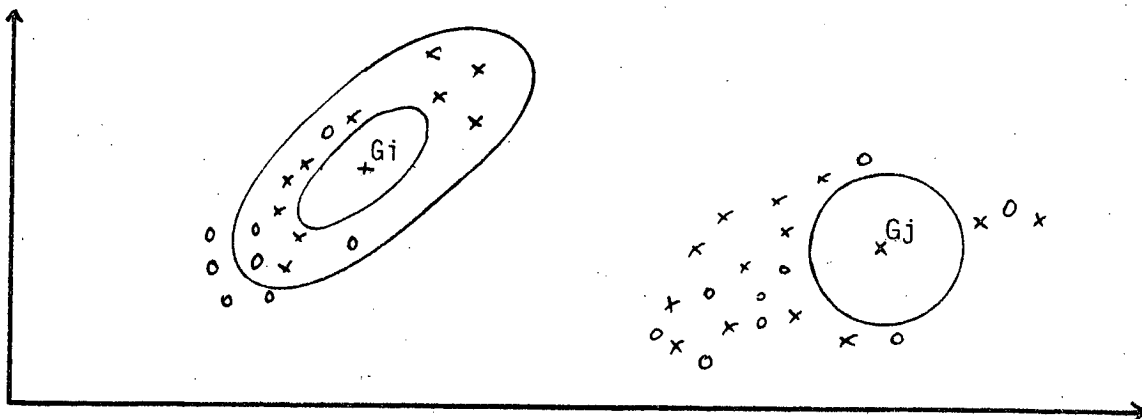
C'est évident : le cas le plus favorable pour rendre le critère meilleur que  $tol$  est, qu'à partir de  $K$ ,  $V_n$  soit rempli qu'avec des éléments de  $im(K)$ . Le critère vaudra alors, pour  $K = k_{\max}$ ,  $(Maj+k_{\max}-K)/k_{\max}$  qui est inférieur à  $tol$  par hypothèse.

### IV-4. QUELQUES REMARQUES SUR LE TRAITEMENT INFORMATIQUE

Informatiquement les étapes 1 et 2 de l'algorithme sont confondues, le test s'effectuant au fur et à mesure du calcul des  $di(X_{\text{new}}, G_i)$ .

Il est possible de modifier légèrement l'algorithme à plusieurs niveaux :

- L'étape 3 peut être simplifiée en ne retenant que les régions telles que  $X_{\text{new}}$  soit contenu dans leur ellipse circonscrite ( $di(X_{\text{new}}, G_i) < R_{\text{MAX}i}$ ). Dans ce cas, le min de l'étape 1 ne sera calculé que sur ces régions.
- On peut également dans un but de simplification (gain de temps et surtout de place) ne pas conserver en mémoire les points d'une région qui appartiennent à son ellipse inscrite. Le fichier d'apprentissage se présente alors comme suit :



Dans ce cas on considère simplement que les éléments intérieurs à une région n'interviennent pas lorsque l'affectation n'est pas directe (on prendra alors sans aucun doute un  $k_{\max}$  plus faible dans la procédure KPPV).

- Programmation de la procédure KPPV( $k_{\min}, k_{\max}, \text{tol}$ ) :

informatiquement la procédure KPPV n'est pas équivalente au calcul des  $K$  plus proches voisins  $k_{\max} - k_{\min} + 1$  fois, bien heureusement! Elle consiste à chaque étape à mettre à jour un tableau appelé VOTE dont la dimension est égale au nombre total de modalités et tel que

$$\text{VOTE}(i) = \text{card} (V(K) \cap A_i)$$

(0) Le critère  $P(K_0)$  est initialisé à la valeur  $\text{tol}$ .

(1) Calcul des  $k_{\min}$  plus proches voisins, initialisation de VOTE, calcul de  $P(k)$ , test.

(2) Pour les autres  $K$  on teste le nouvel arrivant (sa modalité est-elle  $\text{im}(K-1)$  ou non ?) La mise à jour de VOTE,  $\text{Maj}(K)$ ,  $P(K)$ ,  $\text{im}(K)$  en est grandement facilitée.

(3) Les tests ont, bien entendu, lieu au fur et à mesure, les valeurs optimales écrasant les précédentes ; on effectue en premier les tests de sortie définitive ( $P(K) > \max(\text{tol}, k_{\max}/k_{\max}+K)$ ) et  $\text{Maj}+k_{\max}-K/k_{\max} < \text{tol}$ ).

En résumé, lors de la programmation de l'algorithme nous avons essayé de gagner de la place et du temps à tous les niveaux pour rendre la méthode utilisable dans tous les cas de figure, y compris dans le cas d'un très grand volume de données.

#### IV-5. AVANTAGES DE LA PROCEDURE

Le gain de temps obtenu grâce au prétraitement permet d'effectuer le calcul des  $K$  plus proches voisins pour plusieurs valeurs de  $K$ . Ceci permet de résoudre 2 problèmes importants :

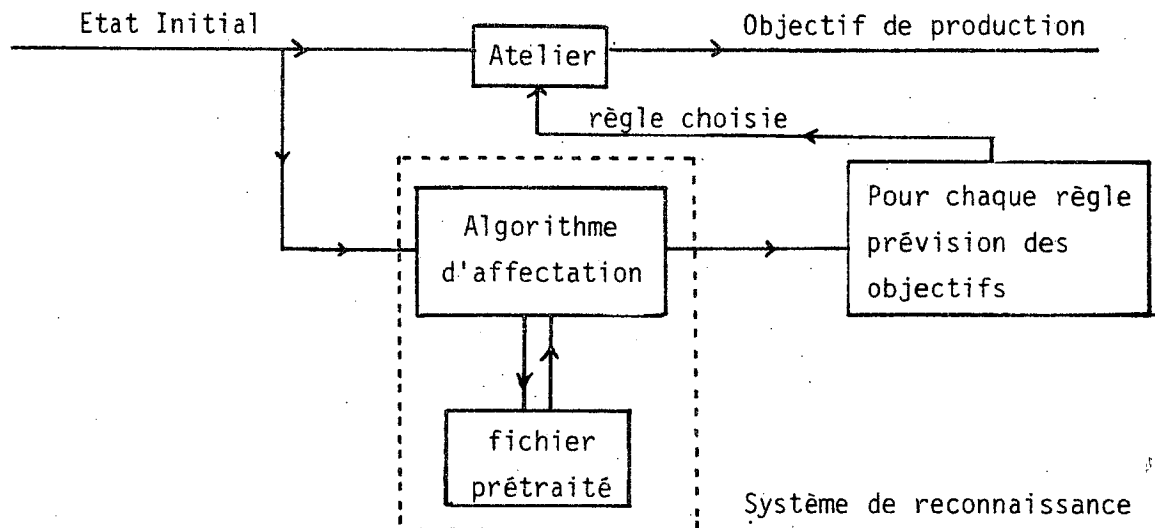
- Le choix du meilleur  $K$ .

- On ne rejette l'affectation P.P.V. qui si l'on est dans un cas d'indétermination pour toute valeur de  $K$ . Par exemple, supposons que pour  $k = 7$  la boule contienne 4 éléments de la classe 1 et 3 éléments de la classe 2. En ne regardant que cette valeur de  $K$  on rejetterait sans doute l'affectation alors qu'il se peut que pour  $k = 4$  tous les éléments appartiennent à la classe 1.

La méthode proposée n'est donc pas seulement un algorithme rapide P.P.V. mais bien une "méthode compromis" entre une méthode géométrique et une méthode de voisinage (utilisation des métriques de Mahalanobis associées aux régions, affectation géométrique lorsque l'affectation P.P.V. n'est pas suffisamment sûre).

## V - APPLICATIONS

Nous avons intégré notre méthode dans un système de reconnaissance pour l'ordonnancement d'ateliers en temps réel qui constitue notre centre d'intérêt principal. Le système se présente comme suit :



Il s'agit donc de prévoir l'objectif qualitatif (respect des délais, diminution des en-cours,...) que l'on obtiendra après application d'une certaine règle de gestion (FIFO, priorité au produit le plus en retard,...).

On dispose ainsi d'un fichier d'apprentissage issu de simulations précisant pour un état initial défini par  $p$  variables (commande, niveau d'en-cours,...) la classe d'objectifs obtenue (par exemple, délais non respectés, diminution des en-cours) après application d'une règle donnée. Lorsqu'un nouvel état se présente, on veut pouvoir dire quelle classe d'objectifs va être atteinte après application de cette même règle.

### V-1. LES DONNEES

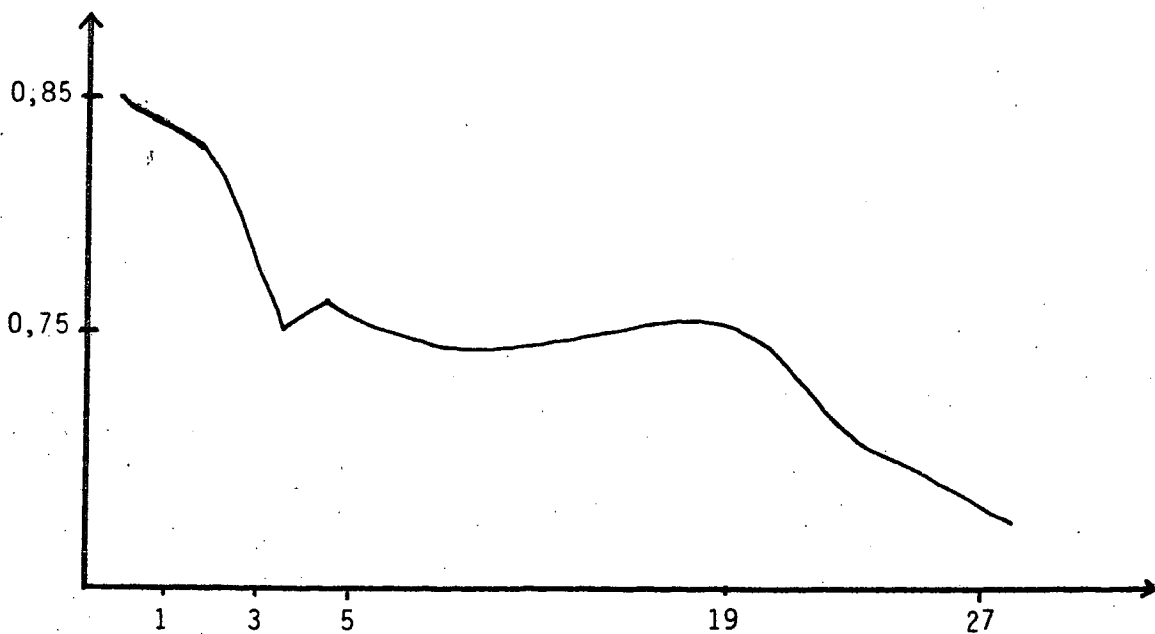
On dispose d'un fichier de 460 éléments définis par deux variables d'état (les quotas de 2 des 3 produits qui vont entrer dans l'atelier) et la classification obtenue après application de la règle F.I.F.O.. Le nombre total de modalités est 8. On décide de conserver 85 % des éléments pour le fichier d'apprentissage et 15 % pour le fichier tests tirés aléatoirement.

### V-2. PRETRAITEMENT

On effectue le prétraitement en région  $(s_0, s_1)$  du fichier d'apprentissage avec les valeurs  $s_0 = 3$ ,  $s_1 = 8$  (pour des valeurs de  $s_1$  inférieures, les matrices de variance-covariance risqueraient d'être singulières). On obtient ainsi 12 régions réparties de la façon suivante (cf. schéma) :

- 2 régions associées à chaque modalité 1,3,6,7
- 1 région associée à chaque modalité 2,4,8

Courbe  $I_s(K)$



### V-3. AFFECTATION

On a ensuite affecté les éléments du fichier test et comparé ces affectations aux appartenances réelles. Les résultats sont les suivants : plusieurs combinaisons de valeurs kmin,kmax,tol ont été essayées. Les résultats dépendent peu de kmin,kmax (i.e. on obtient sensiblement les mêmes résultats pour (2,7),(2,9),(3,11) etc...).

Par contre le seuil de rejet tol semble être optimal entre 2/3 et 3/4 (le pourcentage de réussite chutant très nettement à partir de 3/4).

---

#### AFFECTATION DE 69 NOUVEAUX ELEMENTS PAR LA METHODE MIXTE

---

kmin=2    kmax=9    tol=0.7    metrique=MAHALANOBIS

Affectation directe:35  
Nombre de bien affectes:35

Affectation geometrique apres rejet:2  
Nombre de bien affectes:2

Affectation P.P.V:32  
Nombre de bien affectes:25

POURCENTAGE TOTAL DE REUSSITE:89.855072%

---

kmin=2    kmax=9    tol=0.    metrique=USUELLE

Affectation directe:35  
Nombre de bien affectes:35

Affectation geometrique apres rejet:0  
Nombre de bien affectes:0

Affectation P.P.V:34  
Nombre de bien affectes:25

POURCENTAGE TOTAL DE REUSSITE:86.956522%

---

kmin=2                    kmax=9                    tol=0.8                    metrique=MAHALANOBIS

Affectation directe:35  
Nombre de bien affectes:35

Affectation geometrique apres rejet:8  
Nombre de bien affectes:3

Affectation P.P.V:26  
Nombre de bien affectes:21

POURCENTAGE TOTAL DE REUSSITE:85.507246%

-----  
On a donne ici les meilleurs resultats obtenus (2,9,0.7,Mah), la 2eme serie de resultats soulignant l'importance du seuil de rejet et de la distance de mahalanobis.

Remarque

-----  
Les 69 affectations sont obtenues quasi-instantanement.

#### 5.4 Comparaison avec d'autres methodes

-----  
On donne rapidement les resultats obtenus apres applications de methodes classiques:

#### AFFECTATION DE 69 NOUVEAUX ELEMENTS PAR LA METHODE GEOMETRIQUE

-----  
Nombre d'elements testes:69  
Nombre de bien affectes:51

POURCENTAGE TOTAL DE REUSSITE:73.913043%

#### AFFECTATION DE 69 NOUVEAUX ELEMENTS PAR LA METHODE CLASSIQUE P.P.V

-----  
k=3

Nombre d'elements testes:69  
Nombre de bien affectes:57

POURCENTAGE TOTAL DE REUSSITE:82.608696%

-----  
k=5

Nombre d'elements testes:69



Nombre de bien affectes:56

POURCENTAGE TOTAL DE REUSSITE:81.15942%

---

k=9

Nombre d'elements testes:69  
Nombre de bien affectes:57

POURCENTAGE TOTAL DE REUSSITE:82.608696%

---

## CONCLUSION

Nous pensons que ce que nous venons de décrire permet de traiter de nombreux cas en analyse discriminante prédictive, en particulier lorsqu'une méthode du type plus proches voisins s'avère efficace. Elle permet un gain de temps et de précision.

Deux inconvénients demeurent néanmoins :

- La nécessité d'effectuer un prétraitement parfois très long (surtout si le nombre de variables prédictives est important).
- La place mémoire occupée par la nouvelle structure des données (fichier des régions et de leurs caractéristiques).

Ainsi, on peut n'utiliser, si on le souhaite, qu'une des étapes de la méthode :

- Le prétraitement uniquement afin de déterminer les régions et leur centre de gravité. On appliquera ensuite une méthode géométrique classique, en raisonnant sur les régions et non pas sur les classes.
- L'algorithme d'affectation sans prétraitement en raisonnant sur les classes et non pas sur le partitionnement des régions.

## BIBLIOGRAPHIE

- [ 1 ] : K. FUKUNAGA and P. NARENDRA, "A branch and bound algorithm for computing k-nearest neighbours", IEEE, May 1973.
- [ 2 ] : J.H. FRIEDMAN, "An algorithm for finding nearest neighbours", IEEE, October 1975.
- [ 3 ] : D. SALAS ALVES, "Structures récursives : application à la recherche des plus proches voisins et à la classification", (Thèse).
- [ 4 ] : T. YUNCK, "A technique to identify nearest neighbours", IEEE, Oct 1976.
- [ 5 ] : Pierre A. DEVIJVER, "Reconnaissance des formes par la méthode des plus proches voisins".
- [ 6 ] : J.M. ROMEDER, "Méthodes et programmes d'analyse discriminante", Dunod 1973.
- [ 7 ] : F. BONNEAU, J.M. PROTH, "Applications de règles de gestion à un système de fabrication : classification des objectifs atteints en vue de leur utilisation", Rapport de Recherche INRIA n° 372, mars 1985.
- [ 8 ] : E. DIDAY, "Eléments d'analyse des données".

**Imprimé en France**

**par**

**l'Institut National de Recherche en Informatique et en Automatique**

